

Pendekatan Hybrid Clustering – Classification Untuk Klasifikasi Risiko Kredit Menggunakan K- Means Dan Random Forest

^{1*}Nisaa' Husnia Yahya, ²Danar Putra Pamungkas, ³Made Ayu Dusea
Widyadara

^{1,2,3} Teknik Informatika, Universitas Nusantara PGRI Kediri

E-mail: ¹yahyanisaa461@gmail.com, ²danar@unpkediri.ac.id, ³madedara@unpkediri.ac.id

Penulis Korespondensi : Nisaa' Husnia Yahya

Abstrak— Risiko kredit merupakan salah satu tantangan utama dalam lembaga keuangan karena dapat menyebabkan kerugian finansial dan mengganggu stabilitas operasional. Penelitian ini bertujuan mengklasifikasikan risiko kredit menggunakan pendekatan hybrid yang mengombinasikan K-Means dan Random Forest. Dataset yang digunakan terdiri dari 7.155 data pinjaman dengan atribut besar pinjaman, sisa pinjaman, angsuran, tenor, dan tunggakan maksimum. K-Means digunakan untuk membentuk label risiko secara otomatis berdasarkan pola data, yang selanjutnya dimanfaatkan sebagai target klasifikasi dan fitur tambahan pada Random Forest. Hasil penelitian menunjukkan bahwa model hybrid memiliki performa lebih baik dibandingkan model baseline dengan akurasi 100%, sedangkan model baseline mencapai 98,94%. Analisis feature importance menunjukkan bahwa fitur cluster dan tunggakan maksimum merupakan faktor yang paling berpengaruh dalam menentukan risiko kredit. Temuan ini menunjukkan bahwa pendekatan hybrid efektif meningkatkan kinerja klasifikasi pada dataset tanpa label.

Kata Kunci— Klasifikasi, Clustering, Risiko Kredit, K-Means, Random Forest

Abstract— Credit risk is one of the main challenges facing financial institutions because it can lead to financial losses and disrupt operational stability. This study aims to classify credit risk using a hybrid approach that combines K-Means and Random Forest. The dataset consists of 7,155 loan records with attributes including loan amount, outstanding balance, installment amount, loan term, and maximum arrears. K-Means is used to automatically generate risk labels based on data patterns, which are then utilized as classification targets and additional features in the Random Forest model. The results show that the hybrid model performs better than the baseline model with 100% accuracy, while the baseline model achieves 98.94%. Feature importance analysis indicates that the cluster and maximum arrears features are the most influential factors in determining credit risk. These findings demonstrate that the hybrid approach effectively improves classification performance on unlabeled datasets.

Keywords— Classification, Clustering, Credit Risk, K-Means, Random Forest

This is an open access article under the CC BY-SA License.



I. PENDAHULUAN

Perkembangan teknologi informasi dan peningkatan volume data telah mendorong pemanfaatan teknik *data mining* dalam berbagai sektor, termasuk pada bidang keuangan dan

koperasi simpan pinjam. Salah satu permasalahan utama dalam sektor tersebut adalah risiko kredit, yaitu kemungkinan terjadinya kegagalan pembayaran oleh nasabah atau anggota yang dapat berdampak pada kerugian finansial serta menurunkan stabilitas operasional lembaga keuangan [1]. Oleh karena itu, diperlukan suatu pendekatan yang mampu mengidentifikasi dan mengklasifikasikan tingkat risiko secara akurat guna mendukung proses pengambilan keputusan.

Metode *machine learning* telah banyak digunakan dalam klasifikasi risiko kredit. Salah satu metode yang populer adalah *Random Forest* yang merupakan algoritma berbasis *ensemble learning* dengan kemampuan menghasilkan prediksi yang stabil dan akurat [2]. *Random Forest* bekerja dengan membangun sejumlah pohon keputusan dan menggabungkan hasil prediksinya untuk meningkatkan performa klasifikasi [3]. Namun demikian, metode ini termasuk dalam kategori *supervised learning* yang membutuhkan label sebagai target, sehingga menjadi kendala ketika dataset tidak memiliki label yang tersedia secara langsung.

Untuk mengatasi keterbatasan tersebut, metode *unsupervised learning* seperti K-Means *clustering* dapat digunakan untuk menghasilkan label berdasarkan pola dalam data [4]. K-Means merupakan algoritma yang mengelompokkan data berdasarkan kemiripan karakteristik sehingga mampu mengidentifikasi pola tersembunyi dalam dataset [5]. Dalam konteks risiko kredit, metode ini dapat dimanfaatkan untuk membentuk kategori risiko seperti rendah, sedang, dan tinggi secara otomatis.

Pendekatan *hybrid* yang mengombinasikan metode *clustering* dan klasifikasi telah digunakan untuk meningkatkan performa model dalam proses prediksi [6]. Integrasi antara K-Means dan *Random Forest* memungkinkan pemanfaatan hasil segmentasi data sebagai label sekaligus sebagai fitur tambahan, sehingga model dapat memperoleh informasi yang lebih *representatif* dalam proses klasifikasi.

Tabel 1. Penelitian Terdahulu

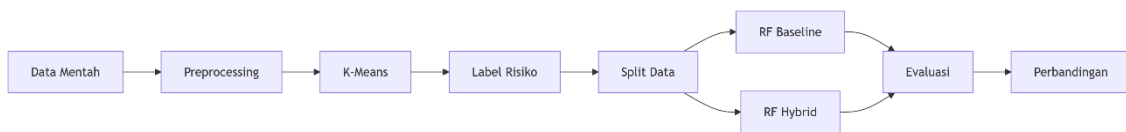
No	Peneliti	Metode	Dataset & tujuan	Hasil
1	Novrian dkk. [7]	K-Means	Data kredit bermasalah Bank Mega Finance (5.305 data)	Menghasilkan 3 cluster risiko (rendah, sedang, tinggi) dengan DBI = 0.201
2	Nur Illah dkk.[8]	K-Means	Data pembayaran kredit nasabah	Berhasil mengelompokkan risiko kredit menjadi 3 kategori secara efektif
3	Ardiansyah & Supatman [9]	K-Means	Data pinjaman online (999 data)	Menghasilkan 2 cluster tagihan (tinggi dan rendah) dengan distribusi 11% dan 89%
4	Damayanti dkk. [10]	Random Forest	Data nasabah lembaga keuangan mikro	Klasifikasi risiko kredit (rendah, sedang, tinggi) dengan akurasi tinggi (Precision 89%)
5	Hanun & Zailani [11]	Random Forest	Data koperasi Mitra Sejahtera	Prediksi kelayakan kredit dengan akurasi 87.88%

Berdasarkan Tabel 1, penelitian terdahulu umumnya menggunakan K-Means untuk *clustering* atau Random Forest untuk klasifikasi secara terpisah. Belum banyak penelitian yang memanfaatkan hasil *clustering* tidak hanya sebagai pembentuk label risiko, tetapi juga sebagai fitur tambahan dalam model klasifikasi. Kondisi ini menunjukkan adanya peluang untuk mengembangkan pendekatan *hybrid* yang lebih efektif.

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan pendekatan hybrid yang mengombinasikan K-Means dan Random Forest untuk menghasilkan label risiko dari data tanpa label serta menganalisis pengaruh penambahan fitur cluster terhadap performa klasifikasi. Evaluasi dilakukan menggunakan accuracy dan confusion matrix untuk membandingkan kinerja model baseline dan model hybrid.

II. METODE

Penelitian ini menggunakan pendekatan data mining dengan metode *machine learning* untuk melakukan klasifikasi risiko kredit. Pendekatan yang digunakan adalah metode *hybrid* yang mengombinasikan algoritma K-Means sebagai metode *clustering* dan *Random Forest* sebagai metode klasifikasi untuk meningkatkan performa prediksi risiko kredit. Alur penelitian ditunjukkan pada Gambar 1 yang menggambarkan tahapan preprocessing, clustering, serta klasifikasi menggunakan dua skenario model, yaitu baseline dan hybrid.



Gambar 1. Alur Penelitian

A. Dataset dan Variabel Penelitian

Dataset yang digunakan merupakan data pinjaman anggota yang terdiri dari beberapa atribut, yaitu Besar Pinjaman, Sisa Pinjaman, Angsuran Pokok, Tenor, dan Tunggalan. Data yang digunakan merupakan data historis pinjaman anggota pada periode 2020 – 2025 dengan jumlah total sebanyak 7.155 data.

Selain atribut asli, dibentuk fitur Tunggalan Maksimum yang digunakan dalam proses clustering dan klasifikasi. Dataset dibagi menjadi data *training* dan data *testing* dengan rasio 80:20 menggunakan teknik *stratified sampling*. Data *training* digunakan untuk melatih model *Random Forest*, sedangkan data *testing* merupakan data yang tidak digunakan dalam proses pelatihan dan digunakan untuk menguji performa model dalam mengklasifikasikan tingkat risiko kredit.

B. Preprocessing Data

Tahap *preprocessing* dilakukan untuk menyiapkan data sebelum proses modeling. Proses ini meliputi pembersihan data, pengecekan *missing value*, serta transformasi atribut. Selain itu, dilakukan pengolahan data dengan mengambil satu data untuk setiap anggota berdasarkan nilai tenor terkecil. Untuk setiap anggota dipilih data dengan tenor terkecil guna merepresentasikan kondisi pinjaman terbaru.

Data kemudian dinormalisasi menggunakan metode Min-Max Scaling untuk menyamakan skala antar fitur. Proses normalisasi dirumuskan sebagai berikut [12]:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Proses ini juga memastikan bahwa setiap anggota direpresentasikan secara unik tanpa redundansi data.

C. Clustering Menggunakan K-Means

K-Means dipilih karena kemampuannya dalam mengelompokkan data numerik secara efisien. Tahap *clustering* dilakukan menggunakan algoritma K-Means untuk mengelompokkan data berdasarkan kemiripan karakteristik. Algoritma ini bekerja dengan meminimalkan jarak antara data dengan centroid cluster yang dinyatakan sebagai [13]:

$$WCSS = \sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \quad (2)$$

Semakin kecil nilai WCSS, Semakin baik kualitas pengelompokan data yang dihasilkan. Penentuan jumlah *cluster* dilakukan menggunakan metode *Elbow* dan divalidasi menggunakan *Silhouette Score* serta *Davies-Bouldin Index* untuk mendapatkan jumlah cluster yang optimal. Model K-Means diinisialisasi dengan jumlah *cluster* sebanyak 3 dan parameter *random_state* untuk memastikan hasil yang konsisten.

D. Pembentukan Label Risiko

Label risiko dibentuk berdasarkan hasil *clustering* dengan mengurutkan nilai rata-rata tunggakan maksimum pada setiap *cluster*. *Cluster* dengan nilai terendah dikategorikan sebagai risiko rendah, sedangkan *cluster* dengan nilai tertinggi dikategorikan sebagai risiko tinggi. Label ini kemudian digunakan sebagai target dalam proses klasifikasi. Pendekatan ini memungkinkan pembentukan label risiko secara otomatis dari data tanpa label (*unsupervised learning*) [14].

E. Klasifikasi Menggunakan Random Forest

Tahap klasifikasi dilakukan menggunakan algoritma *Random Forest* yang merupakan metode *ensemble learning* dengan kemampuan memberikan prediksi yang akurat dan stabil [15]. Model Random Forest diimplementasikan menggunakan parameter yang telah ditentukan untuk menghasilkan model yang stabil dan efisien. Model dilatih menggunakan data *training* dan diuji menggunakan data *testing* untuk menghasilkan prediksi kategori risiko dengan tingkat akurasi dan stabilitas yang tinggi.

F. Model Baseline dan Hybrid

Penelitian ini menggunakan dua skenario model, yaitu model *baseline* dan model *hybrid*. Model *baseline* menggunakan fitur asli tanpa menambahkan informasi *cluster*, sedangkan model *hybrid* menggunakan fitur tambahan berupa *cluster* hasil K-Means. Perbandingan kedua model dilakukan untuk menganalisis pengaruh penggunaan fitur cluster terhadap peningkatan performa klasifikasi.

G. Evaluasi Model

Evaluasi model dilakukan menggunakan metrik *accuracy* dan *confusion matrix*. *Accuracy* digunakan untuk mengukur tingkat ketepatan prediksi model, sedangkan *confusion matrix* digunakan untuk menganalisis distribusi kesalahan klasifikasi.

Selain itu, dilakukan analisis *feature importance* untuk mengidentifikasi fitur yang paling berpengaruh dalam proses klasifikasi. Selain evaluasi pada proses klasifikasi, dilakukan juga pengujian stabilitas *cluster* menggunakan *Adjusted Rand Index* (ARI). Pengujian ini bertujuan untuk mengukur tingkat konsistensi hasil *clustering* K-Means terhadap perubahan nilai *random*

state. Nilai ARI berada pada rentang -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan bahwa hasil *clustering* memiliki tingkat kestabilan yang tinggi.

III. HASIL DAN PEMBAHASAN

A. Dataset

Tabel 2. Contoh Dataset Pinjaman Anggota

ID Anggota	Tanggal Pencairan	Besar Pinjaman	Sisa Pinjaman	Angsuran Pokok	Jasa	Tunggakan	Tenor
722	4/5/2019	2,400,000	200,000	200,000	12,000	270,000	1
509	8/19/2019	1,500,000	250,000	125,000	15,000	270,000	2
696	6/14/2019	1,500,000	375,000	125,000	22,500	148,500	3
702	7/12/2019	3,000,000	1,000,000	250,000	60,000	1,080,000	4
93	11/7/2019	2,040,000	850,000	170,000	51,000	1,953,900	5

Berdasarkan Tabel 2, setiap anggota memiliki karakteristik pinjaman yang berbeda. Atribut Besar Pinjaman menunjukkan jumlah dana yang diterima anggota, di mana nilai pinjaman yang lebih besar dapat meningkatkan risiko apabila tidak diimbangi dengan kemampuan pembayaran yang baik. Sisa Pinjaman menunjukkan kewajiban yang masih harus dibayarkan oleh anggota, sehingga nilai yang tinggi mengindikasikan beban pinjaman yang lebih besar. Angsuran Pokok menggambarkan besarnya cicilan yang harus dibayar pada setiap periode, sedangkan Jasa menunjukkan biaya tambahan atau bunga yang dikenakan pada pinjaman. Atribut Tunggakan merepresentasikan keterlambatan pembayaran yang terjadi, sehingga semakin besar nilai tunggakan maka semakin tinggi potensi risiko kredit. Sementara itu, Tenor menunjukkan jangka waktu pinjaman yang dimiliki anggota. Perbedaan nilai pada setiap atribut tersebut digunakan sebagai dasar dalam proses *clustering* dan klasifikasi risiko kredit.

B. Preprocessing

Tahap *preprocessing* dilakukan untuk menyiapkan data sebelum proses *clustering* dan klasifikasi. Pada tahap ini, dilakukan pemilihan satu data untuk setiap anggota berdasarkan nilai tenor terkecil yang merepresentasikan kondisi pinjaman terakhir. Proses ini dilakukan untuk menghindari duplikasi akibat adanya beberapa transaksi pinjaman pada anggota yang sama, sehingga setiap anggota hanya direpresentasikan oleh satu data. Setelah proses tersebut, jumlah data berkurang dari 7.155 data menjadi 471 data yang sesuai dengan jumlah anggota unik, sehingga redundansi data dapat dikurangi dan representasi data menjadi lebih relevan.

Selain menggunakan atribut asli berupa Besar Pinjaman, Sisa Pinjaman, Angsuran Pokok, Jasa, Tunggakan, dan Tenor, dibentuk pula fitur Tunggakan Maksimum yang diperoleh dari nilai tunggakan terbesar yang pernah dimiliki setiap anggota selama periode pinjaman. Fitur ini digunakan untuk merepresentasikan tingkat keterlambatan pembayaran tertinggi yang pernah terjadi pada masing-masing anggota sehingga dapat memberikan informasi tambahan dalam proses pembentukan cluster dan klasifikasi risiko kredit.

Selanjutnya, dilakukan pengecekan *missing value* dan hasilnya menunjukkan bahwa tidak terdapat data yang hilang dalam dataset. Data kemudian dinormalisasi menggunakan Persamaan (1) sehingga seluruh fitur berada pada rentang nilai 0 hingga 1. Proses normalisasi dilakukan untuk menghindari dominasi fitur yang memiliki skala lebih besar serta meningkatkan kualitas hasil *clustering* dan klasifikasi.

C. Hasil Clustering dan Evaluasi

Proses *clustering* dilakukan menggunakan algoritma K-Means dengan jumlah *cluster* yang ditentukan melalui metode *Elbow* serta divalidasi menggunakan *Silhouette Score* dan *Davies-Bouldin Index* [16]. Penggunaan dua metrik evaluasi bertujuan untuk memastikan kualitas *cluster* dari sisi kepadatan dan pemisahan antar *cluster*.

Hasil evaluasi jumlah *cluster* ditunjukkan pada Tabel 3.

Tabel 3. Evaluasi Jumlah Cluster

K	Silhouette score	Davies-Bouldin index
2	0.563	0.966
3	0.557	0.901
4	0.466	0.847
5	0.430	0.881

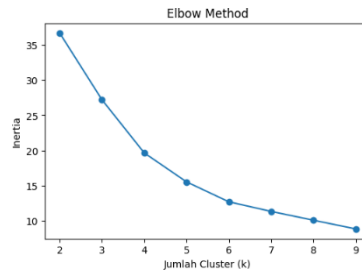
Berdasarkan Tabel 3, nilai *Silhouette Score* tertinggi diperoleh pada $k=2$ sebesar 0,563, yang menunjukkan bahwa anggota dalam *cluster* memiliki tingkat kemiripan yang baik dan pemisahan antar *cluster* cukup jelas. Sementara itu, nilai *Davies-Bouldin Index* terendah diperoleh pada $k=4$ sebesar 0,847, yang mengindikasikan bahwa *cluster* yang terbentuk memiliki tingkat overlap yang lebih kecil. Meskipun demikian, pada $k=3$ diperoleh nilai *Silhouette Score* sebesar 0,557 dan *Davies-Bouldin Index* sebesar 0,901 yang masih menunjukkan kualitas *cluster* yang baik. Selain itu, penggunaan tiga *cluster* menghasilkan pembagian kategori risiko rendah, sedang, dan tinggi yang lebih mudah diinterpretasikan. Oleh karena itu, $k=3$ dipilih sebagai jumlah *cluster* optimal untuk merepresentasikan karakteristik risiko kredit berdasarkan data pinjaman.

Tabel 4. Uji Stabilitas Cluster Menggunakan ARI

Perbandingan Random State	Nilai ARI
1 vs 10	1.0000
10 vs 20	0.9984
20 vs 30	1.0000
30 vs 42	0.9219

Berdasarkan Tabel 4, diperoleh nilai ARI sebesar 1,0000 pada perbandingan random state 1 vs 10 dan 20 vs 30, sedangkan nilai ARI terendah diperoleh pada perbandingan random state 30 vs 42 sebesar 0,9219. Rata-rata nilai ARI yang diperoleh sebesar 0,9801. Nilai ARI yang mendekati 1 menunjukkan bahwa hasil pengelompokan yang dihasilkan dari beberapa nilai *random state* memiliki tingkat kesesuaian yang sangat tinggi. Hal ini menunjukkan bahwa perubahan inisialisasi centroid awal tidak memberikan perbedaan yang signifikan terhadap hasil *clustering*. Dengan demikian, algoritma K-Means menghasilkan *cluster* yang stabil dan pola pengelompokan yang terbentuk relatif konsisten meskipun menggunakan nilai *random state* yang berbeda.

D. Analisis dan Visualisasi Clustering



Gambar 2. Elbow Method

Gambar 2 menunjukkan hasil evaluasi jumlah cluster menggunakan metode Elbow berdasarkan nilai Within Cluster Sum of Squares (WCSS) yang dinyatakan pada Persamaan (2). Nilai WCSS menunjukkan jumlah kuadrat jarak antara setiap data dengan centroid cluster, sehingga semakin kecil nilai WCSS maka semakin homogen anggota dalam setiap cluster. Pada Gambar 2 terlihat bahwa penurunan nilai inertia berlangsung cukup signifikan dari $k=1$ hingga $k=3$, kemudian mulai melandai setelah $k=3$. Kondisi tersebut menunjukkan bahwa penambahan jumlah cluster setelah $k=3$ tidak memberikan penurunan WCSS yang signifikan. Oleh karena itu, $k=3$ dipilih sebagai jumlah cluster optimal. Selain didukung oleh metode Elbow, pemilihan $k=3$ juga mempertimbangkan hasil Silhouette Score dan Davies-Bouldin Index serta menghasilkan pembagian risiko yang lebih mudah diinterpretasikan menjadi risiko rendah, sedang, dan tinggi.

Selanjutnya, hasil *clustering* dianalisis berdasarkan nilai rata-rata atribut pada setiap cluster yang ditunjukkan pada Tabel 5.

Tabel 5. Rata-rata Atribut Setiap Cluster

Cluster	Besar Pinjaman	Sisa Pinjaman	Angsuran Pokok	Tenor	Tunggakan Maksimum
0	3.161.407	753.819	271.595	3	106.352
1	3.099.574	684.681	258.298	2	2.582.368
2	14.315.192	4.161.154	1.059.231	3	869.381

Berdasarkan Tabel 5, Cluster 0 memiliki rata-rata tunggakan maksimum sebesar 106.352 yang merupakan nilai terendah dibandingkan cluster lainnya. Kondisi tersebut menunjukkan bahwa anggota pada cluster ini memiliki riwayat keterlambatan pembayaran yang relatif kecil sehingga dikategorikan sebagai risiko rendah. Cluster 1 memiliki rata-rata tunggakan maksimum sebesar 2.582.368 yang merupakan nilai tertinggi, sehingga anggota pada cluster ini memiliki potensi gagal bayar yang lebih besar dan dikategorikan sebagai risiko tinggi. Sementara itu, Cluster 2 memiliki rata-rata besar pinjaman, sisa pinjaman, dan angsuran pokok yang paling tinggi, dengan nilai tunggakan maksimum sebesar 869.381 yang berada di antara kedua cluster lainnya. Oleh karena itu, cluster ini dikategorikan sebagai risiko sedang. Hasil tersebut menunjukkan bahwa algoritma K-Means mampu memisahkan kelompok anggota berdasarkan karakteristik pinjaman dan tingkat tunggakan secara efektif.

Hasil clustering ini menunjukkan bahwa algoritma K-Means mampu mengelompokkan data berdasarkan pola karakteristik pinjaman dan tingkat tunggakan secara efektif. Informasi cluster yang dihasilkan kemudian digunakan sebagai label risiko sekaligus fitur tambahan dalam proses klasifikasi menggunakan Random Forest.

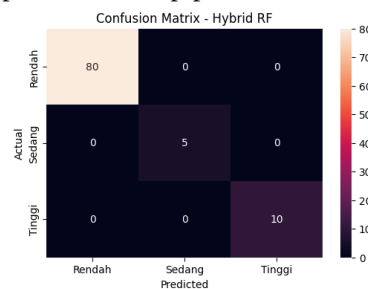
E. Hasil Klasifikasi dan Evaluasi Model

Setelah proses pembentukan label risiko, dilakukan klasifikasi menggunakan algoritma *Random Forest* dengan dua skenario model, yaitu model *baseline* dan model *hybrid*. Hasil evaluasi model ditunjukkan pada Tabel 6.

Tabel 6. Perbandingan Akurasi Model

Model	Akurasi
Baseline	98.94%
Hybrid (dengan cluster)	100%

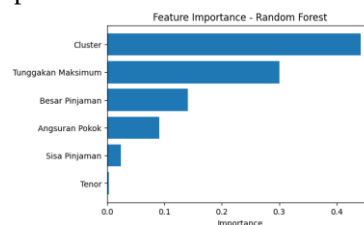
Berdasarkan Tabel 6, model baseline menghasilkan akurasi sebesar 98,94%, sedangkan model hybrid menghasilkan akurasi sebesar 100%. Peningkatan akurasi tersebut menunjukkan bahwa informasi cluster hasil K-Means memberikan tambahan informasi yang mampu membantu *Random Forest* dalam membedakan karakteristik risiko kredit. Dengan demikian, penambahan fitur cluster memberikan kontribusi positif terhadap performa model klasifikasi.



Gambar 4. Confusion Matrix Model Hybrid

Gambar 4 menunjukkan confusion matrix dari model hybrid. Seluruh data pada masing-masing kelas berhasil diklasifikasikan dengan benar, yaitu 80 data pada kelas risiko rendah, 5 data pada kelas risiko sedang, dan 10 data pada kelas risiko tinggi. Tidak terdapat nilai pada elemen di luar diagonal utama confusion matrix, yang menunjukkan bahwa tidak terjadi kesalahan klasifikasi antar kelas. Kondisi tersebut menghasilkan akurasi sebesar 100%, sehingga model hybrid memiliki kemampuan yang sangat baik dalam membedakan tingkat risiko kredit berdasarkan karakteristik data yang digunakan. Meskipun demikian, pengujian pada dataset yang lebih beragam masih diperlukan untuk mengevaluasi kemampuan generalisasi model terhadap data baru.

F. Hasil dan Analisis Feature Importance



Gambar 5. Feature Importance Random Forest

Gambar 5 menunjukkan bahwa fitur Cluster memiliki nilai importance tertinggi dibandingkan fitur lainnya. Hal ini menunjukkan bahwa informasi hasil clustering memberikan kontribusi terbesar dalam proses pengambilan keputusan pada model *Random Forest*. Selain itu, fitur Tunggakan Maksimum juga memiliki pengaruh yang cukup besar, yang menunjukkan bahwa riwayat keterlambatan pembayaran merupakan faktor penting dalam menentukan tingkat risiko

kredit. Sementara itu, fitur Besar Pinjaman, Angsuran Pokok, Sisa Pinjaman, dan Tenor memiliki nilai importance yang lebih rendah sehingga kontribusinya terhadap proses klasifikasi relatif lebih kecil. Hasil ini menunjukkan bahwa kombinasi informasi cluster dan indikator tunggakan berperan penting dalam meningkatkan kemampuan model dalam membedakan tingkat risiko kredit.

IV. KESIMPULAN

Penelitian ini berhasil menerapkan pendekatan *hybrid* yang mengombinasikan metode K-Means dan *Random Forest* dalam klasifikasi risiko kredit. Metode *K-Means* digunakan untuk membentuk label risiko secara otomatis berdasarkan pola data, sedangkan *Random Forest* digunakan untuk melakukan klasifikasi dengan memanfaatkan fitur asli dan fitur tambahan berupa *cluster*.

Hasil penelitian menunjukkan bahwa pendekatan *hybrid* mampu meningkatkan performa model dibandingkan model *baseline*. Model *hybrid* menghasilkan akurasi sebesar 100%, lebih tinggi dibandingkan model *baseline* sebesar 98,94%. Hal ini menunjukkan bahwa penambahan fitur *cluster* memberikan kontribusi signifikan dalam meningkatkan kemampuan model dalam mengklasifikasikan risiko kredit.

Selain itu, hasil analisis menunjukkan bahwa fitur *Cluster* dan Tunggakan Maksimum merupakan faktor utama yang berpengaruh dalam menentukan tingkat risiko kredit. Hasil pengujian menggunakan *Adjusted Rand Index* (ARI) juga menunjukkan bahwa *cluster* yang dihasilkan memiliki tingkat stabilitas yang tinggi terhadap perubahan nilai *random state*, sehingga pendekatan *clustering* yang digunakan dinilai cukup konsisten dan andal. Hal ini menunjukkan bahwa kombinasi antara hasil *clustering* dan indikator keuangan mampu menghasilkan model klasifikasi yang lebih efektif.

Dengan demikian, pendekatan *hybrid K-Means* dan *Random Forest* dapat menjadi solusi yang efektif dalam permasalahan klasifikasi risiko kredit, khususnya pada dataset yang tidak memiliki label. Pendekatan ini juga berpotensi untuk diimplementasikan dalam sistem pendukung keputusan pada lembaga keuangan guna meningkatkan akurasi analisis risiko.

DAFTAR PUSTAKA

- [1] A. A. Montevechi, R. de C. Miranda, A. L. Medeiros, and J. A. B. Montevechi, "Advancing credit risk modelling with Machine Learning: A comprehensive review of the state-of-the-art," *Eng. Appl. Artif. Intell.*, vol. 137, p. 109082, Nov. 2024, doi: 10.1016/J.ENGAPPAI.2024.109082.
- [2] A. Primajaya and B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 1, no. 1, pp. 27–31, Mar. 2018, doi: 10.24014/IJAIDM.V1I1.4903.
- [3] D. Ismafillah *et al.*, "Analisis algoritma pohon keputusan untuk memprediksi penyakit diabetes menggunakan oversampling smote," *INFOTECH: Jurnal Informatika & Teknologi*, vol. 4, no. 1, pp. 27–36, Jun. 2023, doi: 10.37373/INFOTECH.V4I1.452.
- [4] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
- [5] F. P. F. Pangestu, N. Y. N. Yasin, R. C. R. C. Hasugian, and yunita yunita, "Penerapan Algoritma K-Means Untuk Mengklasifikasi Data Obat," *Jurnal Sisfokom (Sistem*

- Informasi dan Komputer*), vol. 12, no. 1, pp. 53–62, Mar. 2023, doi: 10.32736/SISFOKOM.V12I1.1461.
- [6] F. D. Rahman, M. I. Z. Mulki, and A. Taryana, “CLUSTERING DAN KLASIFIKASI DATA CUACA CILACAP DENGAN MENGGUNAKAN METODE K-MEANS DAN RANDOM FOREST,” *Jurnal SINTA: Sistem Informasi dan Teknologi Komputasi*, vol. 1, no. 2, pp. 90–97, Apr. 2024, doi: 10.61124/SINTA.V1I2.15.
- [7] W. Novrian, A. Afriani, and J. Purnama Sari, “Segmentation of Problematic Loan Customers Using the K-Means Clustering Algorithm to Support Strategic Decision-Making (Case Study: Bank Mega Finance Bengkulu),” doi: 10.70656/ijcse.v2i01.499.
- [8] S. Nur Illah, N. Suarna, I. Ali, and D. Solihudin, “Journal of Artificial Intelligence and Engineering Applications K-Means Clustering Method to Make Credit Payment Grouping Efficient,” 2025. [Online]. Available: <https://ioinformatic.org/>
- [9] M. Ardiansyah and Supatman, “Klasterisasi Tagihan Pada Nasabah Pinjaman Online Menggunakan Metode K-Means Clustering,” *Jurnal RESTIKOM: Riset Teknik Informatika dan Komputer*, vol. 6, no. 2, pp. 286–294, Aug. 2024, doi: 10.52005/RESTIKOM.V6I2.318.
- [10] F. Damayanti, A. Budiman, S. Sundari, and T. M. Nainggolan, “Classification of Customer Credit Risk Levels Using the Random Forest Method: A Case Study on Microfinance Institutions,” *Journal of Computer Science Artificial Intelligence and Communications*, vol. 1, no. 2, pp. 52–56, Nov. 2024, doi: 10.64803/JOCAIC.V1I2.20.
- [11] N. Listiana Hanun and A. Udin Zailani, “PENERAPAN ALGORITMA KLASIFIKASI RANDOM FOREST UNTUK PENENTUAN KELAYAKAN PEMBERIAN KREDIT DI KOPERASI MITRA SEJAHTERA,” doi: 10.37365/jti.v6i1.61.
- [12] M. Rizal Ubaidillah, N. Adhi Santoso, E. Ungguk Sedya Utami, and S. YMI Tegal, “Penerapan K-Means Clustering untuk Klasifikasi Jenis Beras Berdasarkan Tren Harga di Pasar Induk Beras Cipinang (PIBC),” *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 3, pp. 4300–4306, Sep. 2025, doi: 10.31004/riggs.v4i3.2594.
- [13] A. Kuraria, N. Jharbade, and M. Soni, “Centroid Selection Process Using WCSS and Elbow Method for K-Mean Clustering Algorithm in Data Mining,” *Int. J. Sci. Res. Sci. Eng. Technol.*, pp. 190–195, Dec. 2018, doi: 10.32628/ijrsrset21841122.
- [14] N. K. Zuhail, D. P. Pamungkas, and R. Wulaningrum, “Klasifikasi Emosi Pada Wajah Dengan Menggunakan K-MEANS Clustering dan KDEP,” *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, vol. 5, no. 1, pp. 243–248, Aug. 2021, doi: 10.29407/INOTEK.V5I1.955.
- [15] T. Y. Wen and S. A. Mohd Aris, “Hybrid Approach of EEG Stress Level Classification Using K-Means Clustering and Support Vector Machine,” *IEEE Access*, vol. 10, pp. 18370–18379, 2022, doi: 10.1109/ACCESS.2022.3148380.
- [16] R. D. Nugraha, D. D. Adelia, and D. Rivaldi, “Segmentasi Pelanggan Retail Berbasis Perilaku menggunakan Algoritma K-Means Clustering,” *Digital Transformation Technology*, vol. 5, no. 2, pp. 141–148, Oct. 2025, doi: 10.47709/DIGITECH.V5I2.6340.